

# LIGAND: database of chemical compounds and reactions in biological pathways

Susumu Goto\*, Yasushi Okuno, Masahiro Hattori, Takaaki Nishioka<sup>1</sup> and Minoru Kanehisa

Bioinformatics Center, Institute for Chemical Research, Kyoto University, Uji, Kyoto 611-0011, Japan and

<sup>1</sup>Graduate School of Agricultural Sciences, Kyoto University, Sakyo-ku, Kyoto 606-8502, Japan

Received September 19, 2001; Accepted September 26, 2001

## ABSTRACT

**LIGAND is a composite database comprising three sections: COMPOUND for the information about metabolites and other chemical compounds, REACTION for the collection of substrate–product relations representing metabolic and other reactions, and ENZYME for the information about enzyme molecules. The current release (as of September 7, 2001) includes 7298 compounds, 5166 reactions and 3829 enzymes. In addition to the keyword search provided by the DBGET/LinkDB system, a substructure search to the COMPOUND and REACTION sections is now available through the World Wide Web (<http://www.genome.ad.jp/ligand/>). LIGAND may be also downloaded by anonymous FTP (<ftp://ftp.genome.ad.jp/pub/kegg/ligand/>).**

## INTRODUCTION

The completion of the human genome sequence and those of many other organisms, including several dozens of bacteria, accelerated post-genome projects aimed at elucidating the blueprint of life from a scientific point of view. They are also aimed at discovering new drugs and other useful materials, and at deriving biodegradation pathways of xenobiotic chemicals such as pollutants and toxins from medical, industrial and environmental viewpoints. All of them require chemical information, which is not stored in the genome, in addition to information about genes and proteins, which is derived from the genome, and chembioinformatics has been considered as one of the important research fields in the post-genome era. The LIGAND database (1) has been organized to fill in the gap between genomic information and chemical information, and applied to actual reconstruction of metabolic pathways in the completely sequenced organisms in the Kyoto Encyclopedia of Genes and Genomes (KEGG) (2,3).

The LIGAND database is a composite database comprising three sections: COMPOUND, for the information about metabolites and other chemical compounds; REACTION, for the collection of substrate–product relationships representing metabolic and other reactions; and ENZYME, for the information about enzyme molecules. We report here the current status of the LIGAND database, where efforts are being made to add more data in the COMPOUND and REACTION sections, and the new features of the two sections including the substructure search facility.

## CURRENT STATUS OF LIGAND

The COMPOUND and ENZYME sections are constructed as flat-file databases and the data format of each section is similar to those of GenBank (4) flat files: a fixed number of columns is assigned to specify each field of entry (1). COMPOUND and REACTION sections are now organized and maintained as ISIS format (see below).

The COMPOUND section contains a collection of chemical compounds that are found in the KEGG/PATHWAY database and in the ENZYME section, as well as other compounds found in literature. The REACTION section is a collection of chemical reactions, mostly enzymatic reactions, represented as conversions of chemical structures. The ENZYME section is based on the enzyme nomenclature of the International Union of Biochemistry and Molecular Biology (IUBMB) (5), which is also available from the World Wide Web (<http://www.chem.qmw.ac.uk/iubmb/enzyme/>). We have added several links to other databases such as OMIM (6) for human genetic diseases, PROSITE (7) for amino acid sequence motifs, PDB (8) for protein structures, in addition to PATHWAY and GENES databases in KEGG.

The number of entries in the current release is summarized in Table 1.

## NEW FEATURES OF COMPOUND AND REACTION

### New types of compounds: drugs and xenobiotic chemicals

The COMPOUND section was originally created by extracting chemical compounds from the metabolic pathways of the KEGG/PATHWAY database, as well as the ENZYME section of LIGAND. The current version of COMPOUND includes xenobiotic chemicals such as environmental pollutants and toxins, because KEGG has an agreement with UM-BBD (9) to include biodegradation pathways of xenobiotic chemicals in KEGG/PATHWAY. Efforts have also been made to add more drug-related chemicals, and their ratio is increasing. They will be used, for example, as the starting compounds to search possible degradation pathways, which will connect to the existing pathways presented in the KEGG/PATHWAY database.

### Compounds and reactions in the ISIS database

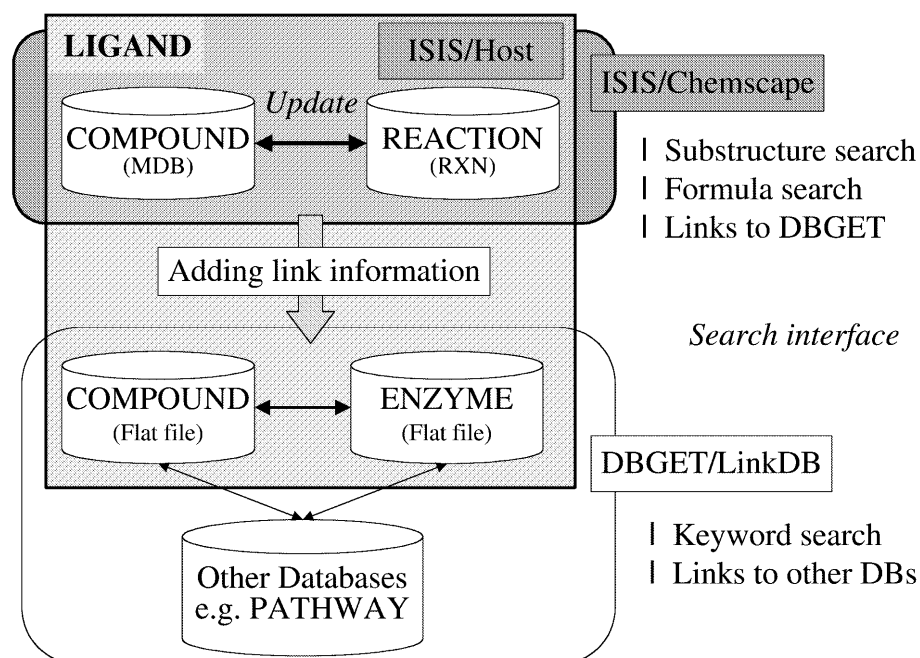
The COMPOUND and REACTION sections are now managed as MDB and RXN formats, respectively, which are database formats for handling chemical structures by the ISIS/HOST database (Fig. 1, upper part). ISIS automatically computes

\*To whom correspondence should be addressed. Tel: +81 774 38 3271; Fax: +81 774 38 3269; Email: goto@kuicr.kyoto-u.ac.jp

**Table 1.** The number of entries in release 20.0 (October 2001) of the LIGAND database

Section	Content	Number
COMPOUND	Entries	7298
	Entries with chemical formulae	6406
	Entries with molecular structures	6002
	Links to ENZYME	4590
	Links to ENZYME as reactants	4426
	Links to ENZYME as cofactors	82
	Links to ENZYME as inhibitors	155
	Links to ENZYME as effectors	33
	Links to CAS	3020
REACTION	Entries	5166
	Reactions defined in ENZYME	4509
	Reactions with known enzymes in KEGG/PATHWAY	2801
	Reactions with unknown enzymes in KEGG/PATHWAY	324
	Non-enzymatic reactions in KEGG/PATHWAY <sup>a</sup>	373
ENZYME	Entries	3829
	Entries with reactions in chemical equations	2906
	Links to KEGG/PATHWAY (metabolic pathways)	1811
	Links to KEGG/GENES (gene catalogs)	1349
	Links to OMIM (human genetic disorders) (6)	440
	Links to PROSITE (proteins sequence motifs) (7)	977

<sup>a</sup>Non-enzymatic reactions include reactions where it is not known whether enzymes are involved in catalysis.



**Figure 1.** Relationship between the ISIS version and the DBGET version of LIGAND. COMPOUND and REACTION are managed by the ISIS and substructure and formula searches are executed to this version. They are exported to the flat file formats where several links to the other databases are added to COMPOUND, and COMPOUND is integrated into the DBGET system, in addition to ENZYME.

molecular formulae and weights once the curator of the COMPOUND database inputs a chemical structure. Curators of REACTION update the information about the substrates and

products by the compound IDs, not by their structures, because we developed a program which automatically imports the compound structures of each reaction from the COMPOUND section.

### Substructure searches using ISIS database and Chemscape

Because COMPOUND and REACTION are stored in the ISIS/HOST database, they can be accessed through the Chemscape server. This enables users to search by compound structures, in addition to the keyword search originally provided by the DBGET/LinkDB system (10,11). Although the Chime plug-in and ISIS/Draw are required for the structure search, they are freely available from the MDL web site (<http://www.mdli.com/>) for academic users. The relationship between the ISIS version and the DBGET version of LIGAND is summarized in Figure 1.

### AVAILABILITY

The LIGAND database is accessible through the World Wide Web at <http://www.genome.ad.jp/ligand/>. The user can then invoke the DBGET/LinkDB system to retrieve COMPOUND and ENZYME, the ISIS/Chemscape-based system to retrieve COMPOUND and REACTION by substructure or chemical formula.

The LIGAND database can be downloaded via anonymous FTP at <ftp://ftp.genome.ad.jp/pub/kegg/ligand/>. This directory contains all sections, COMPOUND, ENZYME and REACTION, including GIF image files and MDL-MOL files for compound structures. The same data set is mirrored at the NCBI repository <ftp://ncbi.nlm.nih.gov/repository/LIGAND/>.

The basic concept of the LIGAND database has been published elsewhere (1). The present article reflects the most up-to-date version of the database and should be cited accordingly.

### ACKNOWLEDGEMENTS

We thank Nobue Takeuchi, Tomoko Komeno, Rumiko Yamamoto and Yuriko Matsuura for inputting the compound and reaction data. We also thank Koichiro Tonomura for

developing the interface of the ISIS system for searching and updating COMPOUND and REACTION. The computational resource was provided by the Bioinformatics Center, Institute for Chemical Research, Kyoto University. This work was supported by the grants from the Ministry of Education, Culture, Sports, Science and Technology of Japan, Japan Society for the Promotion of Science, and Japan Science and Technology Corporation.

### REFERENCES

1. Goto, S., Nishioka, T. and Kanehisa, M. (1998) LIGAND: chemical database for enzyme reactions. *Bioinformatics*, **14**, 591–599.
2. Kanehisa, M., Goto, S., Kawashima, S. and Nakaya, A. (2002) The KEGG databases at GenomeNet. *Nucleic Acids Res.*, **30**, 42–46.
3. Kanehisa, M. (1997) A database for post-genome analysis. *Trends Genet.*, **13**, 375–376.
4. Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Rapp, B.A. and Wheeler, D.L. (2002) GenBank. *Nucleic Acids Res.*, **30**, 17–20.
5. International Union of Biochemistry and Molecular Biology (1992) *Enzyme Nomenclature: Recommendations (1992) of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology*. Academic Press, New York.
6. Pearson, P., Francomano, C., Foster, P., Bocchini, C., Li, P. and McKusick, V. (1994) The status of online Mendelian inheritance in man (OMIM) medio 1994. *Nucleic Acids Res.*, **22**, 3470–3473. Updated article in this issue: *Nucleic Acids Res.* (2002), **30**, 52–55.
7. Falquet, L., Pagni, M., Bucher, P., Hulo, N., Sigrist, C.J.A., Hofmann, K. and Bairoch, A. (2002) The PROSITE database, its status in 2002. *Nucleic Acids Res.*, **30**, 235–238.
8. Westbrook, J., Feng, Z., Jain, S., Bhat, T.N., Thanki, N., Ravichandran, V., Gilliland, G.L., Bluhm, W., Weissig, H., Greer, D.S., Bowne, P.E., Berman, H.M. (2002) The Protein Data Bank: unifying the archive. *Nucleic Acids Res.*, **30**, 245–248.
9. Ellis, L.B.M., Hershberger, C.D., Bryan, E.M. and Wackett, L.P. (2001) The University of Minnesota Biocatalysis/Biodegradation Database: emphasizing enzymes. *Nucleic Acids Res.*, **29**, 340–343.
10. Fujibuchi, W., Goto, S., Migimatsu, H., Uchiyama, I., Ogiwara, A., Akiyama, Y. and Kanehisa, M. (1998) DBGET/LinkDB: an integrated database retrieval system. *Pac. Symp. Biocomput.*, 683–694.
11. Kanehisa, M. (1997) Linking databases and organisms: GenomeNet resources in Japan. *Trends Biochem. Sci.*, **22**, 442–444.